Carbohydrate Research, 212 (1991) 253–259
Elsevier Science Publishers B.V., Amsterdam

Note

A database of three-dimensional structures of monosaccharides from molecular-mechanics calculations

Serge Pérez and Marie-Madeleine Delage
Laboratoire de Physicochimie des Macromolécules, Institut de la Recherche Agronomique, B.P. 527, 44026 Nantes cédex (France)

(Received November 17th, 1989; accepted for publication, in revised form, October 30th, 1990)

Databases for carbohydrates[1,2] are less developed than those for nucleic acids and proteins and include only the primary structure, one or two key references, the source of the carbohydrate, and an index or accession number.

Three-dimensional information on carbohydrates is critical to understanding of interactions between such molecules and other carbohydrates or proteins. Three-dimensional information is also important for the synthesis of analogues of biologically active carbohydrates and for interpreting experimental and modeling data. A prerequisite for understanding the structures of the larger molecules is a detailed description of the structures of their monosaccharide components.

The geometry of crystalline monosaccharides may be determined by crystallographic studies; these structures may show considerable variations arising from such forces as hydrogen bonding and crystal packing and from configurational variation at the glycosidic linkage[3,5]. Numerous crystallographic studies are needed to allow generalizations to be made. Inadvertent selection of an inappropriate geometry may invalidate conformational energy calculations[6] and predictions of spatial positioning of subsequent residues in multi-residue molecules[5,7].

Alternatively, an average geometry may be used, as pioneered by Arnott and Scott[8]. They determined average geometry for both α- and β-D-glucopyranose, based on approximately 20 X-ray and neutron-diffraction structures. Similarly, Sheldrick and Akrigg[9] proposed ideal geometries for eight hexopyranoses, based on 161 examples of pyranose rings in the crystallographic literature; however, the positions of the hydroxyl substituents were not reported.

A third approach, used here, employs a molecular-mechanics computer program considered to give accurate structures and energies for molecules. Molecular mechanics is empirical in that it is developed by fitting equations and parameters to experimental data. Provided that these equations and parameters are applied to unknowns that are similar to the structures used to parametrize the program, the approach thus interpolates known data, offering a measure of confidence in the results.

One of the most elaborate molecular-mechanics programs, used in many fields of

0008-6215/91/$ 03.50

organic chemistry, is Allinger's MM2 program[10]. Here we use MM2CARB, a version[6] that modifies the parameters of the acetal segment to account for stereoelectronic (exo)anomeric effects. MM2CARB correctly reflects the energies and variations in bond-lengths, and bond-angles for various conformers of dimethoxymethane and 2-methoxytetrahydropyran. MM2CARB also adequately reproduces the dependence of ring geometry on the conformation of the glycosidic linkage as observed in the solid state. The MM2CARB method has been used by Tvaroská and Gadjos[11] to generate geometric parameters of ring substituents for both the anomers of the anomeric D-aldohexopyranoses in the $^4C_1$ conformation.

Here, we aim to establish optimized geometries and conformations for common monosaccharides, organized in a standardized, database format.

METHODS

Pyranoid sugars are the commonest monosaccharide units in polysaccharides, glycoproteins and glycolipids. Furanoid sugars occur in DNA and RNA and sometimes in plant and microbial polysaccharides. Septanoid and open-chain sugars are much less common.

The 34 pyranoid monosaccharides listed here (Table I) constitute >95% of the monosaccharides occurring in complex carbohydrates, including all N- and O-linked glycans. Sulphated mononosaccharides and monosaccharides having the idose configuration have been omitted; the latter display conformational variability[12]. The optimizations were all performed on the α- and β-glycosides (Table I).

The conformations of monosaccharides are influenced by such variables as the torsional angles between the hydroxymethyl and hydroxyl groups, and are susceptible to influence from the neighbouring monosaccharide residue. These variations cannot be addressed in the present database. The minimization was therefore performed from one stable orientation as the starting point. A major deficiency of MM2CARB is that it finds only local minima and not the global minimum. The dispositions of rotatable substituents affect the calculated energy values. Primary alcohol groups usually adopt staggered dispositions (*gg*, *gt*, and *tg*) and correspond to local minima[13]. Most molecular-mechanics studies fail to indicate that one orientation of the primary alcohol group has an energy prohibitively higher than the others. Being aware of these pitfalls[4], we ascertained that the initial substituent orientations had no significant effect on the geometry of the monosaccharide until after full minimization.

When available, the starting models came from X-ray or neutron crystallography data[15–29] (Table I). For the optimized monosaccharides, the coordinates of the atomic positions (Å) are presented in conventional orthogonal system (Table II). In calculating the Cartesian coordinates, the *x* axis was in the direction of the C-1–O-5 linkage, and the *x,y* plane was determined by C-1–O-5–C-5. The main advantage of orthogonal Cartesian coordinates is that geometrical properties may be readily calculated.

TABLE I

Monosaccharides included in the database

Name	*Abbreviation*	*Ref.*
3,6-Anhydro-α-D-galactopyranose	3,6An-α-D-Gal*p*	15
3,6-Dideoxy-α-D-*xylo*-hexopyranose (abequose)	α-D-Abe*p*	
α-D-Galactopyranose	α-D-Gal*p*	16
6-*O*-Methyl-α-D-galactopyranose	α-D-Gal*p*6Me	
α-D-Galactopyranuronic acid	α-D-Gal*p*A	
6-*O*-Methyl-α-D-galactopyranosiduronate	α-D-Gal*p*A6Me	
2-Amino-2-deoxy-α-D-galactopyranose	α-D-Gal*p*N	
2-Acetamido-2-deoxy-α-D-galactopyranose	α-D-Gal*p*NAc	17
α-D-Glucopyranuronic acid	α-D-Glc*p*A	
α-D-Glucopyranose	α-D-Glc*p*	18
α-D-Mannopyranose	α-D-Man*p*	18
α-D-Mannopyranuronic acid	α-D-Man*p*A	
N-Acetyl-α-D-neuraminic acid	α-D-Neu*p*5AC	
α-D-Xylopyranose	α-D-Xyl*p*	19
3,6-Anhydro-α-L-galactopyranose	3,6An-α-L-Gal*p*	
α-L-Arabinopyranose	α-L-Ara*p*	20
α-L-Fucopyranose	α-L-Fuc*p*	21
α-L-Galactopyranose	α-L-Gal*p*	
α-L-Gulopyranuronic acid	α-L-Gul*p*A	22
α-L-Rhamnopyranose	α-L-Rha*p*	23
β-D-Arabinopyranose	β-D-Ara*p*	
2-Amino-2-deoxy-β-D-galactopyranose	β-D-Gal*p*N	24
β-D-Galactopyranose	β-D-Gal*p*	16
6-*O*-Acetyl-β-D-galactopyranose	β-D-Gal*p*6Ac	25
6-*O*-Methyl-β-D-galactopyranosiduronate	β-D-Gal*p*A6Me	
β-D-Glucopyranuronic acid	β-D-Glc*p*A	
2-Amino-2-deoxy-β-D-glucopyranose	β-D-Glc*p*N	
β-D-Glucopyranose	β-D-Glc*p*	26
6-*O*-Acetyl-β-D-glucopyranose	β-D-Glc*p*6Ac	27
2-Acetamido-2-deoxy-β-D-glucopyranose	β-D-Glc*p*NAc	28
β-D-Mannuronic acid	β-D-Man*p*A	
β-D-Mannopyranose	β-D-Man*p*	28
β-D-Xylopyranose	β-D-Xyl*p*	29
β-L-Arabinopyranose	β-L-Ara*p*	20

THE DATABASE

The database is a collection of individual files that contain the coordinate entries, identified by the IUPAC name of the monosaccharide.

The coordinate entries constitute the central information store. The atom file contains the following:

A sequence number, indicating the position of an atom in the atom list, to which may be added a residue identifier.

A name for each atom, consisting of 1 to 8 digit characters.

Coordinates: three floating-point numbers indicating the atom's position in space with respect to the origin of the coordinate frame.

TABLE II

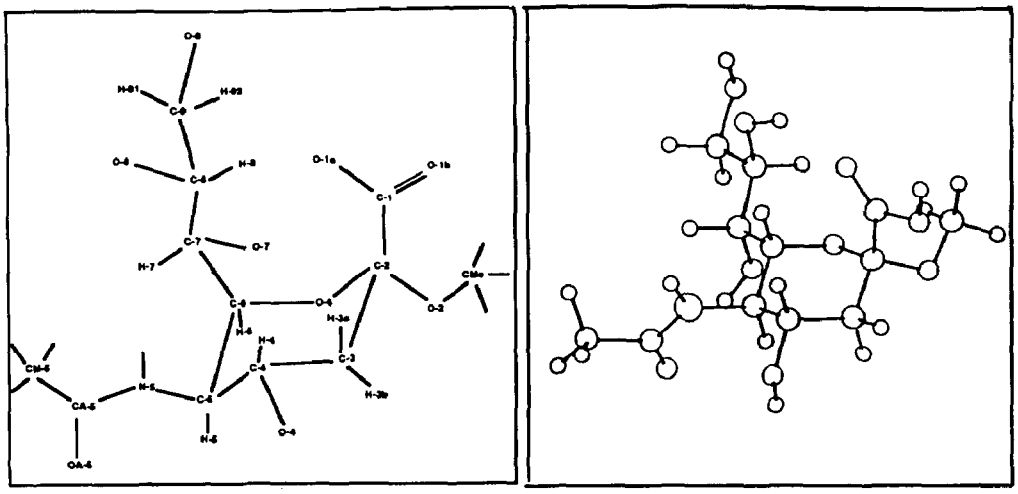
Cartesian coordinates of methyl *N*-acetyl-α-neuraminopyranosidic acid (Me α-D-Neu*p*5Ac)

C-1	−0.5665	−0.7742	1.1992
C-2	0.0000	0.0000	0.0000
C-3	−0.4166	−0.6688	−1.3168
C-4	−1.8727	−0.3935	−1.6689
C-5	−2.0692	1.1196	−1.7094
C-6	−1.7089	1.6809	−0.3328
C-7	−1.8958	3.2021	−0.2489
C-8	−1.6136	3.7780	1.1422
C-9	−1.9260	5.2698	1.2228
CA-5	−3.7798	2.3210	−3.0694
CM-5	−5.2622	2.5139	−3.3290
O-1a	−1.3473	−0.3516	2.0070
O-1b	−0.0483	−2.0156	1.2877
O-2	1.4098	0.0000	0.0000
O-4	−2.2056	−0.9803	−2.9304
O-6	−0.3607	1.3784	0.0000
O-7	−1.0347	3.8362	−1.2025
O-8	−2.4277	3.1066	2.1114
O-9	−1.6196	5.7617	2.5303
OA-5	−2.9412	2.9283	−3.7110
N-5	−3.4795	1.4287	−2.0576
H-3a	−0.2307	−1.7680	−1.2860
H-3b	0.2190	−0.2774	−2.1468
H-4	−2.5452	−0.8610	−0.9108
H-5	−1.3755	1.5417	−2.4739
H-6	−2.3820	1.2076	0.4176
H-7	−2.9440	3.4669	−0.5224
H-8	−0.5464	3.6074	1.4199
H-91	−1.3105	5.8572	0.5040
H-92	−3.0040	5.4689	1.0220
HCM-51	−5.7647	2.9051	−2.4150
HCM-52	−5.4317	3.2413	−4.1554
HCM-53	−5.7318	1.5459	−3.6172
HN-5	−4.2540	0.9956	−1.5513
HO-1b	−0.4219	−2.4387	2.0788
HO-4	−1.7405	−0.5374	−3.6208
HO-7	−1.4720	3.9059	−2.0347
HO-8	−1.9796	2.3308	2.4061
HO-9	−2.1916	5.3350	3.1462
HO-2[a]	1.8057	0.4085	0.7521
CMe[b]	2.0052	0.6143	1.1308
HCMe1[b]	1.6728	0.1248	2.0729
HCMe2[b]	1.7803	1.7041	1.1540
HCMe3[b]	3.1077	0.4902	1.0409

[a] For the reducing form in the α configuration. [b] For the glycosidic methyl group.

Atomic number: the Z number of the atom.

Z typ: an arbitrary number indicating the bonding for atoms in various chemical environments.



```
C-1 _ C-2 = 1.536
C-2 _ C-3 = 1.534
C-3 _ C-4 = 1.523
C-4 _ C-5 = 1.526
C-5 _ C-6 = 1.530
C-6 _ C-7 = 1.535
C-7 _ C-8 = 1.532
C-8 _ C-9 = 1.526
C-1 _ O-1a= 1.200
C-1 _ O-1b= 1.348
C-2 _ O-6 = 1.425
C-2 _ O-2 = 1.410
C-4 _ O-4 = 1.431
C-5 _ N-5 = 1.485
N-5 _ CA-5= 1.382
CA-5_ CM-5= 1.517
CA-5_ OA-5= 1.218
C-6 _ O-6 = 1.421
C-7 _ O-7 = 1.433
C-8 _ O-8 = 1.433
C-9 _ O-9 = 1.430

C-H = 1.11
O-H = 0.97
N-H = 1.02
```

```
C-2 _ C-1 _ O-1a= 126.0
C-2 _ C-1 _ O-1b= 111.9
O-1a_ C-1 _ O-1b= 122.0
C-1 _ C-2 _ C-3 = 110.5
C-1 _ C-2 _ O-2 = 111.7
C-1 _ C-2 _ O-6 = 113.2
C-3 _ C-2 _ O-2 = 105.8
C-3 _ C-2 _ O-6 = 110.7
O-2 _ C-2 _ O-6 = 104.7
C-2 _ C-3 _ C-4 = 112.3
C-3 _ C-4 _ C-5 = 108.0
C-3 _ C-4 _ O-4 = 110.6
C-5 _ C-4 _ O-4 = 110.7
C-4 _ C-5 _ C-6 = 108.0
C-4 _ C-5 _ N-5 = 109.6
C-6 _ C-5 _ N-5 = 111.0
C-5 _ C-6 _ C-7 = 112.6
C-5 _ C-6 _ O-6 = 110.9
C-7 _ C-6 _ O-6 = 108.3
C-6 _ C-7 _ C-8 = 113.6
C-6 _ C-7 _ O-7 = 109.2
C-8 _ C-7 _ O-7 = 109.1
C-7 _ C-8 _ C-9 = 112.2
C-7 _ C-8 _ O-8 = 109.5
C-9 _ C-8 _ O-8 = 107.8
C-8 _ C-9 _ O-9 = 109.9
CM-5_ CA-5_ OA-5= 121.3
CM-5_ CA-5_ N-5 = 114.8
OA-5_ CA-5_ N-5 = 123.9
C-2 _ O-6 _ C-6 = 116.5
C-5 _ N-5 _ CA-5= 120.8
```

```
O-1a_ C-1 _ C-2 _ O-2 = 125.2
O-1a_ C-1 _ C-2 _ C-3 =-117.4
O-1a_ C-1 _ C-2 _ O-6 =   7.4
O-1b_ C-1 _ C-2 _ O-2 = -52.7
O-1b_ C-1 _ C-2 _ C-3 =  64.7
O-1b_ C-1 _ C-2 _ O-6 =-170.5
C-1 _ C-2 _ C-3 _ C-4 =  75.7
O-2 _ C-2 _ C-3 _ C-4 =-163.3
O-6 _ C-2 _ C-3 _ C-4 = -50.5
C-1 _ C-2 _ O-6 _ C-6 = -73.4
C-3 _ C-2 _ O-6 _ C-6 =  51.3
O-2 _ C-2 _ O-6 _ C-6 = 164.8
C-2 _ C-3 _ C-4 _ C-5 =  56.1
C-2 _ C-3 _ C-4 _ O-4 = 177.3
C-3 _ C-4 _ C-5 _ C-6 = -59.3
C-3 _ C-4 _ C-5 _ N-5 = 179.7
O-4 _ C-4 _ C-5 _ N-5 =  58.5
C-4 _ C-5 _ C-6 _ C-7 =-178.9
C-4 _ C-5 _ C-6 _ O-6 =  59.6
N-5 _ C-5 _ C-6 _ C-7 = -58.8
N-5 _ C-5 _ C-6 _ O-6 = 179.8
C-4 _ C-5 _ N-5 _ CA-5=-129.6
C-6 _ C-5 _ N-5 _ CA-5= 111.2
C-5 _ C-6 _ C-7 _ C-8 = 176.9
C-5 _ C-6 _ C-7 _ O-7 = -61.1
O-6 _ C-6 _ C-7 _ C-8 = -60.2
O-6 _ C-6 _ C-7 _ O-7 =  61.8
```

Fig. 1. Example of one card image of the data-bank illustrated for an *N*-acetyl-α-D-neurominic acid residue [methyl α-*N*-acetylneuraminopyranosidic acid (Me α-D-Neu*p*5Ac)].

Van der Waals radius: an intermolecular radius corresponding to the size of atoms in space-filling molecular models.

Covalent radius: an intramolecular radius (A) taken from standard compilations. These radii are useful for searching for bonded atoms.

Colour code: an arbitrary number used to designate atoms by colour for plotting.

Three other items (skip code, type-specification, and arbitrary type number) are for internal use and handling of the data.

This atom list has been designed to cope with future developments, and not all of the cited items are given. From the present atom list, other files may be created that correspond to such established formats as the Cambridge Data Base or Protein Data Base. Conversion into formats corresponding to the requirement of most of the commercial molecular modelling packages has been made. The database does not yet supply any management program for use with the files.

As a complement to the computerized 3-dimensional database, printed information has been generated in the form of card images. Each monosaccharide-unit card image shows a schematic description with the labelling of the atoms, a short version of the atom list containing the atom name and the coordinates, a 3-dimensional representation of the molecule in the "ball and stick" style with perspective, and the bond distances, bond angles, and torsional angles of interest. Figure 1, associated with Table II, displays a typical card image corresponding to the *N*-acetyl-α-neuraminic acid residue.

ACKNOWLEDGMENTS

The authors thank Dr. R. Scaringe (Eastman Kodak Laboratories, Rochester, U.S.A.) for suggesting the atom list format. Appreciation is extended to Dr. A. D. French, (USDA, New Orleans, U.S.A.), Dr. Anne Imberty (INRA, Nantes, France), and Dr. W. Mackie (University of Leeds, U.K.) for their critical reading of the manuscript.

REFERENCES

1 Carbohydrate Structure Data Base, (1987) U.S. Department of Energy/ER-0310, Washington, D.C.
2 S. Doubet, K. Bock, D. Smith, A. Darvill, and P. Albersheim, *TIBS*, 14 (1989) 475–477.
3 N. V. Joshi and V. S. R. Rao, *Biopolymers*, 18 (1979) 2993–3004.
4 A. D. French and V. G. Murphy, *Carbohydr. Res.*, 27 (1973) 391–406.
5 A. D. French and V. G. Murphy, *Polymer*, 18 (1977) 489–494.
6 I. Tvaroska and S. Perez, *Carbohydr. Res.*, 149 (1986) 389–410.
7 R. Lemieux and K. Bock, *Arch. Biochem. Biophys.*, 211 (1983) 125–34.
8 S. Arnott and S. E. Scott, *J. Chem. Soc. Perkin Trans. 2* (1972) 324–335.
9 S. Sheldrick and D. Akrigg, *Acta Crystallogr.*, B36 (1980) 1615–1621.
10 N. L. Allinger, *J. Am. Chem. Soc.*, 99 (1977) 8127–8134.
11 I. Tvaroska and J. Gadjos, *Chem. Pap.*, 41 (1987) 485–500.
12 M. Ragazzi, D. F. Ferro, and A. Provalosi, *J. Comp. Chem.*, 7 (1987) 105–112.
13 R. H. Marchessault and S. Perez, *Biopolymers*, 18 (1979) 2369–2374.
14 A. D. French, V. Tran, and S. Perez, in A. D. French and J. W. Brady (Eds.), *Computer Modeling of Carbohydrate Molecules, ACS Symp. Ser. 30*, 1990, pp. 191–212.

15 D. Lamba, A. L. Segre, S. Glover, W. Mackie, B. Sheldrick, and S. Perez, *Carbohydr. Res.*, 208 (1990) 215–230.
16 S. Takagi and G. A. Jeffrey, *Acta Crystallogr.*, B35 (1979) 902–906.
17 A. Neuman, H. Gillier-Pandraud, F. Longchambon, and D. Rabinovich, *Acta Crystallogr.*, B31 (1975) 474–477.
18 G. A. Jeffrey, R. K. McMullan, and S. Takagi, *Acta Crystallogr.*, B33 (1977) 728–737.
19 S. Takagi and G. A. Jeffrey, *Acta Crystallogr.*, B34 (1978) 3104–3107.
20 S. Takagi and G. A. Jeffrey, *Acta Crystallogr.*, B34 (1978) 1591–1596.
21 F. Longchambon, J. Ohannessian, D. Avenel, and A. Neuman, *Acta Crystallogr.* B31 (1975) 2623–2627.
22 F. Mo, T. J. Brobak, and I. R. Siddiqui, *Carbohydr. Res.*, 145 (1985) 13–24.
23 S. Takagi and G. A. Jeffrey, *Acta Crystallogr.*, B34 (1978) 2551–2555.
24 M. Takai, S. Watanabe, T. Ashida, and M. Kakudo, *Acta Crystallogr.*, B28 (1972) 2370–2376.
25 K. B. Lindberg, *Acta Crystallogr.*, B32 (1976) 645–647.
26 G. A. Jeffrey and S. Takagi, *Acta Crystallogr.*, B33 (1977) 38–742.
27 K. B. Lindberg, *Acta Crystallogr.*, B32 (1976) 642–645.
28 V. Warin, F. Baert, R. Fournet, G. Strecker, G. Spik, B. Fournet, and J. Montreuil, *Carbohydr. Res.*, 76 (1979) 19–22.
29 S. Takagi and G. A. Jeffrey, *Acta Crystallogr.*, B33 (1977) 3033–3040.